wrangle_report

Introduction

In this paper we will describe our wrangling effort made in the section of wrangling weRateDog

project

Data wrangling consists of:

• Gathering data

• Assessing data

• Cleaning data

1.1 Gathering

Gathering Data for this Project composed from three pieces of data as described below:

• The WeRateDogs Twitter archive. We manually downloaded this file manually by clicking

the following link: twitter_archive_enhanced.csv

• The tweet image predictions, i.e., what breed of dog (or other object, animal,

etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) hosted on Udacity's servers and we downloaded it programmatically using python Requests library on the following (URL of the file:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv)

• entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data stored in a line.

1.2 Assessing

After gathering each of the above pieces of data, assess them visually and programmatically for

quality and tidiness issues was our next step. We could detect and document the following quality

issues and tidiness issues.

1.2.1 Quality

Tweets with no images

(Remove rows where there are no images (expanded_urls))

Dataset contains retweets

(Remove retweets)

Contents of 'text' cutoff

(Display full content of 'text' column)

Missing values in 'name' and dog stages showing as 'None'

(Change missing values in 'name' from 'None' to NaN (dog stages already covered)

Tweet ID# 810984652412424192 doesn't contain a rating

(Remove tweet without rating)

Extra characters after '&'

(Remove extra characters after '&' in archive_clean['text'])

Erroneous datatypes (timestamp, source, dog stages, tweet_id, in_reply_to_status_id, in_reply_to_user_id)

(Change datatypes of timestamp to datetime, and dog_stage to categorical ,tweet_id, in_reply_to_status_id, and in_reply_to_user_id to strings)

The source column is not clean

(Clean the content of source column, make it more readable)

It is possible that there are high ratings, but they are really not high

(Calculate the division ratio between the numerator and the denominator to extract reasonable values for evaluations)

1.2.2 Tidiness

Dog "stage" variable in four columns: doggo, floofer, pupper, puppo

(Create dog stage variable and remove individual dog stage columns)

Join 'tweet_info' and 'image_predictions' to 'twitter_archive'

(Add tweet and image to archive table)

1.3 Cleaning

Cleaning our data is the third step in data wrangling. It is where we fixed the quality and tidiness

issues that we identified in the assess step.

We used the two types of cleaning, the manual and programmatic.

2 Conclusion

Data wrangling indeed is a core skill that everyone who works with data should be familiar with

since so much of the world's data isn't clean. If we analyze, visualize, or model our data before we

wrangle it, our consequences could be making mistakes, missing out on cool insights, and wasting

time.